一种改进过采样算法在类别不平衡信用评分中的应用 *

邵良杉，周 玉

(辽宁工程技术大学 系统工程研究所，辽宁 葫芦岛 125105)

摘 要：针对信贷行业信用评分业务中存在的样本类别不平衡问题，首先在信用评分各影响因素 Fisher 比率值分析的基础上确定主要评判指标；而后以基于支持度的过采样算法（SDSMOTE）为样例合成算法，支持向量机（SVM）为基预测器，Boosting 算法为框架构建基于 Fisher-SDSMOTE-ESBoostSVM 的类别不平衡信用评分预测模型；并在基分类器训练结束后引入"淘汰策略"，删除未被正确分类的合成样例，重新生成正类样例并修正样例权重；最后以 UCI 数据库中德国信用数据集为实验样本，F-measure 值和 G-mean 值为评价指标，对比分析 Fisher-SDSMOTE-ESBoostSVM 与其他集成学习算法的预测结果。实验结果表明，Fisher-SDSMOTE-ESBoostSVM 算法应用到信贷行业客户信用评分预测中具有可行性和适应性，且预测准确率较高，具有一定的实际应用价值。

关键词：信用评分；类别不平衡；SDSMOTE 算法；Fisher 准则；支持向量机；集成学习

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2017.12.0798

Application of improved oversampling algorithm in class-imbalance credit scoring

Shao Liangshan, Zhou Yu

(*System Engineering Institute Liaoning Technical University*, *Huludao Liaoning* 125105, *China*)

Abstract: In view of class-imbalance in real credit scoring business of credit industry, firstly, determining the main evaluation indicators of credit scoring based on a comprehensive analysis of the influence factors' Fisher ratio value. Then, choosing the SMOTE based on support degree (SDSMOTE) oversampling algorithm to synthesize new samples, SVM played as the base predictor and Boosting algorithm as the framework, a credit scoring prediction model which associated class-imbalance with Fisher-SDSMOTE-ESBoostSVM theory was proposed. Besides, the "elimination strategy" was introduced to delete the synthetic sample which was not classified accurately, after that synthesized the new positive class sample again and modified the sample weight. Finally, the German credit dataset in the UCI database was selected as the experimental dataset, and F-measure value and G-mean value as evaluation standard, comparing and analyzing the prediction result of Fisher-SDSMOTE-ESBoostSVM model and others ensemble learning algorithm. Experimental results show that the application of Fisher-SDSMOTE-ESBoostSVM algorithm to customer credit score prediction is feasible and applicable, and show a high level of accuracy, which proved that the algorithm have a certain practical application value.

Key words: credit scoring; class-imbalance; SDSMOTE algorithm; Fisher criterion; support vector machine; ensemble learning

0 引言

信用评分模型是一种以客户的信用历史资料为依据，为保障各类金融行业的金融安全、运用定量统计分析方法而设定的一种评估或预测信用风险的划定模型。近年来信贷行业规模和涉及领域不断扩大使得信用评分问题日益突出，如何建立高效、可靠的信用评分模型显得尤为重要。

目前，已有一些学者将基于统计和机器学习的方法应用到信用评分模型构建中，如逻辑回归[1]、支持向量机[2]、提升树[3]等方法，并取得了较好的效果。但应用信用评分模型解决实际问题的过程中存在一些不容忽视的问题，如前期筛选使得"好"客户数量较"坏"客户多，将"好"客户错分为"坏"客户与将"坏"客户错分为"好"客户的代价是不同的，信用评分模型所涉及评判指标的维度较高、数据之间存在冗余等。因此构建信用评分模型是一种类别不平衡、数据间冗余较高的学习问题。目前，采样和代价敏感学习是处理类不平衡问题的常用方法。代价敏感学习要求明确错分的代价，而信贷业务中较难准确评估错分的代价，在实际问题中更多地采用采样的方法。采样方法分为过采样和欠采样方法两种，随机欠采样主要是随机删除负类（多数类）中部分样例，对正类（少数类）没有采取

收稿日期：2017-12-07；**修回日期：**2018-01-22 **基金项目：**国家自然科学基金资助项目（71371091）；辽宁省社会规划项目（L14BTJ004）

作者简介：邵良杉（1961-），男，辽宁凌源人，教授，博导，博士，主要研究方向为矿业系统工程、数据挖掘（m18304232428@163.com）；周玉（1994-），女，辽宁鞍山人，硕士研究生，主要研究方向为数据挖掘.

任何操作，Herrera [4] 提出一种基于 K-nearest neighbor 的有指导的欠采样方法，通过保留正类附近的负类样本有效避免关键信息丢失；Blake 等人[5]提出 Balance Cascade 算法，对负类样例不重复采样、固定正类样例，最终建立多个子分类器形成联合分类器。但欠采样方法在删除负类样例的过程中，难免会删除部分含有效信息的负类样例。随机过采样通过随机复制正类样例改善了类间不平衡度，但易出现过拟合问题，对此学者们提出了不同的改进方法，Chawla 等人[6]提出一种新的过采样方法 SMOTE(synthetic minority over-sampling technique)，通过在正类样例及其临近正类样例连线上随机选取一点合成正类样例来解决数据失衡问题； Han 等人[7]在 SMOTE 算法的基础上提出了 Borderline -SMOTE 算法，通过在边缘区域内进行插值使新生成样本更加有效；Nakamura 等人[8]提出基于密度的 SMOTE 改进算法，根据正类样本的分类密度形成聚类簇来控制新样本的合成。

在现有研究的基础上，本文提出一种基于支持度的改进过采样 SMOTE 算法—SDSMOTE（SMOTE based on support degree，SDSMOTE）来处理客户信用评分问题中类别不平衡问题，而后以 Boosting 集成学习方法为框架，SVM 为基学习器，迭代过程中引入“淘汰策略”（elimination strategy），删除被基分类器错误分类的正类合成样本来确保合成样本的质量；此外，鉴于信用评分问题涉及的评判指标维数较高，在合成正类样例前根据各指标 Fisher 比率值来筛选指标。

1 理论分析

1.1 Boost-SVM 基本原理

支持向量机（support vector machine，SVM）[9,10]核心思想是建立一个分类超平面作为决策曲面，最大化正负类之间的隔离边缘。SVM 首先设定训练集 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}$，其中 $x_i\in X$ 为特征向量，$y_i\in Y=\{+1,-1\}$ $(i=1,2,\cdots,l)$；选取适当的核函数 $K(x_1,x_2)$ 和参数 C，构造求解最优化问题：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{j}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_jK(x_i,x_j)-\sum_{j=1}^{l}\alpha_j\\ &s.t.\quad \sum_{i=1}^{l}y_i\alpha_i=0,\quad 0\le\alpha_i\le C,i=1,\cdots,l\end{aligned}\tag{1}$$

得到最优解：$\alpha^*=(\alpha_1^*,\cdots\alpha_l^*)^T$。选取 α^* 的一个正分量 $0<\alpha_j^*<C$，并据此计算阈值 $b^*=y_i-\sum_{i=1}^{l}y_i\alpha_i^*K(x_i-x_j)$；构造决策函数：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^*y_iK(x_1,x_2)+b^*)。\tag{2}$$

Boosting 算法与 SVM 有一个共同点，即在学习过程中注重“最富信息”的样本点[11,12]。Boosting 算法在初次训练时赋予每个样本相同的概率，进入迭代后，对分类错误的样本加大权重使得在下一次迭代中可以更加关注这些点。Boost-SVM 算法拟将支持向量机作为集成学习机框架的学习器，即以支持向量机为 Boosting 算法的基分类器，进一步提高学习机的泛化能力。

1.2 改进的 SMOTE 算法

SMOTE [13,14]基本思想是通过线性内插法在两个临近的正类样本中合成新的正类样本，其可有效避免过度拟合问题，但其无法指导如何选取正类样本及合成新样本。针对以上不足，本文提出一种基于支持度的 SDSMOTE 算法。SDSMOTE 算法通过计算各正类样本的支持度来确定边界样本，可以实现有选择、有差别地合成边界样本的目标，提高合成样本的质量。具体方法如下：

a)使用 Tomek links 数据清理技术对样本数据集中噪声点进行清除。

b)设去除噪声后数据集中正类样本数为 m，负类样本数为 n，样本维数为 d。随机选取一个正类样本 $X_i=(x_{i1},x_{i2},\cdots,x_{it})$ $(i=1,2,\cdots,m)$，利用式（3）计算到 X_i 到每一个负类样本 $y_j(j=1,2,\cdots,n)$ 之间的欧式距离和。

$$S_i=\sum_{j=1}^{n}\sqrt{(x_{i1}-y_{j1})^2+\cdots+(x_{id}-y_{id})^2}\tag{3}$$

c)计算所有 S_i 加和，并以此根据式（4）计算得到正负类样本之间的平均距离：

$$S_{ave}=\sum_{i=1}^{m}S_i\Big/(m\times n)\tag{4}$$

d)将 S_{ave} 设置为距离参数，选取一个正类样本 X_i 为圆心并以距离参数为半径画一个圆，计算每个圆区域内负类样本个数作为正类样本的支持度 k_i，支持度较大意味着正类样本 X_i 被分配一个较高的选择可能性值 $P_i=k_i\Big/\sum_{i=1}^{m}k_i$；相反，样本会被分配一个较小 P_i。

e)设定需要合成的正类样本数 L 为数据集中正类样本与负类样本的差值，根据正类样本的 P_i 值，可以得到以每个正类样本附近需要合成的新样本个数 $l_i=P_i\cdot L$，设需要合成的正类样本为 $X_{new}=\{x_{new1},x_{new2},\cdots,x_{newd}\}$，对被选的正类样本使用改进的差值公式：

$$X_{new}=X_i+rand(0,1)\times(X_{max}-X_i)\tag{5}$$

其中：$X_{\max}$ 为以 X_i 为圆心的圆中距离 X_i 最远的正类样本点，添加新合成的样本到数据集中参与训练和测试。

1.3 基于 Fisher 准则的特征选择

Fisher 准则[15]是一种基于距离的特征选择方法之一，其基本思想是鉴别性能较强的特征，即表现为类内距离尽可能小，类间距离尽可能大的特征。采用单个特征的 Fisher 比值作为准则，并以此对特征进行排序，可以选出鉴别性能较强的特征，从而达到降维的目的。在特征选择过程中，设定存在训练样本 $\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}$，其中 n 为样本数量，$x_i\in R^k$，k 为特征向量的维数，$y_i=\{-1,1\}^l$ 为类别标号，1 表示正类，-1 表

示负类。每一类中包含 n_i 个样本，$x_i^{(k)}$，$m_i^{(k)}$，$m^{(k)}$ 分别为第 i 类样本 x_i，第 i 类样本的均值，所有样本的均值在第 k 维上的取值。用 $S_B^{(k)}$ 和 $S_\omega^{(k)}$ 表示该维特征在训练样本集上的类间方差和类内方差。

$$S_B^{(k)}=\sum_{i=1}^{2}\frac{n_i}{n}\left(m_i^{(k)}-m^{(k)}\right)^2 \tag{6}$$

$$S_\omega^{(k)}=\frac{1}{n}\sum_{i=1}^{2}\sum_{x\in\omega_i}(x^{(k)}-m_i^{(k)})^2=\frac{1}{n}(\delta_1^2+\delta_2^2) \tag{7}$$

则单个特征的 Fisher 准则比值可以表示为

$$J_{fisher}(k)=S_B^{(k)}/S_\omega^{(k)} \tag{8}$$

将 J_{fisher} 称为特征的 Fisher 比率值，某维特征在训练集上的 Fisher 比率值越大，说明该维特征的类别区分度越好，包含越多的鉴别信息，噪声特征的 J_{fisher} 值趋近于 0。

2　仿真实验及性能分析

2.1　Fisher-SDSMOTE-ESBoostSVM 算法实现

本文首先根据 Fisher 比率值对各评判指标进行选择，而后借助过采样 SDSMOTE 算法合成正类样本，分类算法选取 Boosting 算法为框架并以 SVM 模型为基分类器，同时引入“淘汰策略”删除被基分类器错误分类的合成正类样例，最终构建基于 Fisher-SDSMOTE-ESBoostSVM 的类不平衡数据集分类模型。其具体实现过程如下：

a）输入样本数据 Z，借助 Fisher 准则方法在样本集中进行权重计算，并输出特征权重向量 W。根据特征权重值对属性进行筛选，构成降维后新的样本集 S。

b）输入新样本集，应用 SDSMOTE 算法合成少数类样本。将合成样本添加到样本集 S 中。

c）对新样本集中的每个样例设置相同的初始权重值。

d）调用 SVM 学习算法，形成基分类器，借助 Boosting 权重更新过程使下一次迭代时，被当前基分类器错分的样例可以得到更多关注；同时引入“淘汰策略”，删除错误合成的正类样例，消除其对集成学习器分类效果的影响。

e）根据正类样例减少个数重新执行 SDSMOTE 算法合成新样本，将合成样例添加实验样本集后重新规范化权值。权值计算公式如下：

$$w_t^{new}(i)=\begin{cases}\dfrac{1}{n_t}, & x_i\in S_t\\[2mm] w_t^{old}(i)\times\dfrac{n_t-m}{n_t}, & x_i\notin S_t\end{cases} \tag{9}$$

其中：w_t^{old}、w_t^{new} 分别表示第 t 次迭代时，合成样例和原始样例权值；n_t 为训练样本个数；m 为淘汰的正类合成样本个数；S_t 为第 t 次迭代重新合成的正类样例集合。

f）重复执行 d）e）步，直到达到迭代次数 T，计算各基分类器权重，最后组合基分类器形成强分类器。具体流程如图 1 所示。

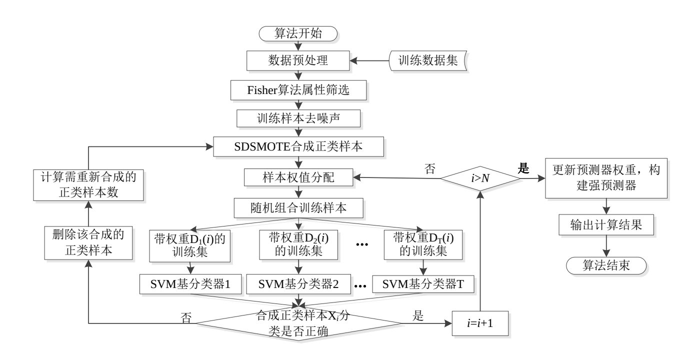


图 1　算法实现流程

2.2　评价指标

单纯地将准确率（accuracy）作为模型分类效果的评价机制对不平衡数据来说意义不大，不少学者针对不平衡数据预测提出了一些更加合理的评价机制[16]，如特异性（specificity））、敏感性（sensitivity）、正类的查准率（precision）、几何平均值（G-mean）、正类的 F-measure 值等。两类别情况下，正类和负类的预测情况可具体分为 TP（实际正类，预测正类）、FP (实际负类，预测正类)、FN(实际负类，预测正类)、TN(实际负类，预测负类)四种。定义各度量计算公式为

特异性 $$Specificity=\frac{TN}{TN+FP}$$

敏感性 $$Sensitivity=\frac{TP}{TP+FN}$$

正类查准率 $$Precision=\frac{TP}{TP+FP}$$

几何平均值 $$G-mean=\sqrt{Sensitivity\cdot Specificity}$$

正类 F-measure $$F-measure=\frac{(1+\beta^2)Sensitivity\cdot Precision}{\beta^2 Sensitivity+Precision}$$

在仅考虑模型正类预测性能的情况下，敏感性 sensitivity 和正类查准率 precision 是相对重要的度量，正类 F-measure 值是敏感性和查准率的调和均值，其计算结果接近两者中的较小者，故较大的 F-measure 值对应的 sensitivity 和 precision 较大，其中 β 通常取值为 1。需要同时考虑模型对两类的预测性能，即希望 TP 和 TN 都较大时，可以使用 G-mean 衡量模型在两个类别上的平均性能。因此，本文选取正类 F-measure 和 G-mean 作为模型分类效果的评价指标。

2.3　算法有效性验证

为测试本文所建模型对类不平衡数据集的分类效果，选取了来自 UCI 数据库和 KEEL 数据库中 5 组不同的数据集进行实验，各数据集的特征信息如表 1 所示。实验过程中采用 10 折交叉验证（10-fold cross-validation）的测试方法，将数据集等分为 10 份，轮流选择其中 9 份作为训练集，1 份作为测试集。对 10 次实验结果中各度量值取均值作为模型的最终评价结果。经反复测试实验参数设置为：Boosting 迭代次数为 500 次，基分类器分数为 20 个，SVM 中核函数选择径向基（RBF）函数，核参数 γ 取值为 2，C 取值为 100。

表 1 不平衡数据集特征与分布

数据集	特征	正类	负类	不平衡比	约简特征
Pima	8	268	500	0.538	7
Ionosphere	34	126	225	0.563	29
wpbc	35	46	148	0.316	31
Wine	13	48	178	0.266	11
Sonar	60	77	168	0.449	52

为同时验证所提出的 Fisher-SDSMOTE-ESBoostSVM 算法中 SDSMOTE 的性能、特征提取及“淘汰策略”的有效性。实验另外分别测试不经特征提取使用的 SMOTE- BoostSVM、SDSMOTE-BoostSVM 算法，采用特征提取的 Fisher-SMOTE-BoostSVM 、Fisher-SDSMOTE-BoostSVM 四种算法所得出的 specificity 、sensitivity、F-measure 和 G- mean 值。表 1 最后一栏中列出了 Fisher 准则提取的特征数情况。表 2 为以上五种算法的各评价指标值的对比情况。

分析表 2 可以看出，经特征提取的两类算法和不经过特征提取的两类算法的比较中，显然经过特征提取的算法分类效果较好，说明 Fisher 算法有效地提取出了关键属性，剔除了不相关或冗余的特征，达到了提高模型精确度，减少运行时间的目的。而 Fisher-SDSMOTE-BoostSVM 的分类效果要远好于 Fisher-SMOTE-BoostSVM 算法，说明基于 SDSMOTE 算法通过有选择、有差别地合成边界样本目标，在一定程度上有效避免了 SMOTE 合成新样本的盲目性，提高了正类合成样本的质量进而提高正类样本的分类准确率。Fisher- SDSMOTE-ESBoostSVM 算法分类效果更优于 Fisher- SDSMOTE-BoostSVM 算法具有大幅度提升，表明结合“淘汰策略”的 SDSMOTE-BoostSVM 算法具有更好的分类性能。综合以上实验结果表明，本文所构建的 Fisher- SDSMOTE-ESBoostSVM 分类器模型，在不同空间结构以及不同维度的不平衡数据集下拥有更强的正负类识别率、更好的综合性能。

表 2 5 种不平衡数据集的评价机制数值

数据集	算法	Specificity	Sensitivity	G-mean	F-measure
Pima	SMOTE-BoostSVM	0.704 ± 0.43	0.559 ± 0.133	0.627 ± 0.036	0.642 ± 0.042
	Fisher-SMOTE-BoostSVM	0.778 ± 0.21	0.611 ± 0.104	0.687 ± 0.030	0.698 ± 0.035
	SDSMOTE- BoostSVM	0.740 ± 0.52	0.724 ± 0.039	0.732 ± 0.025	0.764 ± 0.037
	Fisher-SDSMOTE-BoostSVM	0.81 ± 0.034	0.798 ± 0.24	0.804 ± 0.013	0.823 ± 0.028
	本文算法	0.959 ± 0.017	0.893 ± 0.016	0.925 ± 0.009	0.894 ± 0.006
Ionosphere	SMOTE- BoostSVM	0.741 ± 0.149	0.55 ± 0.109	0.638 ± 0.059	0.658 ± 0.102
	Fisher-SMOTE- BoostSVM	0.765 ± 0.103	0.625 ± 0.082	0.691 ± 0.083	0.735 ± 0.087
	SDSMOTE- BoostSVM	0.79 ± 0.078	0.692 ± 0.091	0.739 ± 0.071	0.781 ± 0.074
	Fisher-SDSMOTE- BoostSVM	0.823 ± 0.066	0.796 ± 0.054	0.809 ± 0.052	0.847 ± 0.067
	本文算法	0.881 ± 0.050	0.962 ± 0.031	0.921 ± 0.039	0.884 ± 0.044
wpbc	SMOTE- BoostSVM	0.692 ± 0.119	0.604 ± 0.086	0.647 ± 0.087	0.654 ± 0.074
	Fisher-SMOTE- BoostSVM	0.797 ± 0.082	0.682 ± 0.073	0.737 ± 0.065	0.773 ± 0.053
	SDSMOTE- BoostSVM	0.826 ± 0.082	0.753 ± 0.076	0.788 ± 0.076	0.766 ± 0.062
	Fisher-SDSMOTE- BoostSVM	0.857 ± 0.073	0.785 ± 0.062	0.820 ± 0.062	0.833 ± 0.047
	本文算法	0.907 ± 0.056	0.894 ± 0.044	0.901 ± 0.037	0.880 ± 0.021
Wine (3vs other)	SMOTE- BoostSVM	0.692 ± 0.111	0.604 ± 0.097	0.647 ± 0.089	0.647 ± 0.102
	Fisher-SMOTE- BoostSVM	0.763 ± 0.089	0.745 ± 0.064	0.754 ± 0.076	0.742 ± 0.082
	SDSMOTE- BoostSVM	0.802 ± 0.091	0.812 ± 0.063	0.807 ± 0.065	0.809 ± 0.073
	Fisher-SDSMOTE- BoostSVM	0.843 ± 0.074	0.832 ± 0.045	0.837 ± 0.051	0.855 ± 0.057
	本文算法	0.923 ± 0.032	0.909 ± 0.036	0.916 ± 0.033	0.925 ± 0.028
Sonar	SMOTE- BoostSVM	0.702 ± 0.109	0.572 ± 0.098	0.634 ± 0.122	0.658 ± 0.097
	Fisher-SMOTE- BoostSVM	0.786 ± 0.082	0.595 ± 0.076	0.684 ± 0.082	0.736 ± 0.067
	SDSMOTE- BoostSVM	0.825 ± 0.087	0.727 ± 0.069	0.774 ± 0.065	0.798 ± 0.059
	Fisher-SDSMOTE-BoostSVM	0.866 ± 0.056	0.776 ± 0.102	0.820 ± 0.058	0.845 ± 0.065
	本文算法	0.947 ± 0.047	0.891 ± 0.049	0.919 ± 0.043	0.914 ± 0.023

2.4 Fisher-SDSMOTE-ESBoosting 算法鲁棒性对比分析

为检验所构建模型的鲁棒性，采用参考文献[17]中的鲁棒性评价机制对以上 5 种算法的鲁棒性进行对比分析，将算法 m 在某一特定数据集上的相对性能用该算法在求解问题时得到的

Adjusted Rand Index 的值与最大 Adjusted Rand Index 值的比值表示。具体计算方法为

$$b_m = R_m / max(R_m) \qquad m = 1,2,\cdots,k \tag{10}$$

在某个数据集上表现最好的算法 m^* 对应的性能 b_{m^*} 为 1，而其他算法的相对性能 b_{m^*} 小于 1，且 b_{m^*} 值越大，相应算法 m 在所有算法中的相对性能越好。因此，本文选取各算法在所有数据集上的 b_{m^*} 值的总和来评价其鲁棒性，总和值越大算法的鲁棒性越强。同样选择以上 5 个数据集为测试数据，将正类与负类样例按 1：10 的比例进行选取，并借助 10 折交叉验证法进行测试。各算法参数设置同上。图 2 为 5 种算法的 G-mean 评价指标的鲁棒性对比（限于篇幅对各算法进行缩写）。

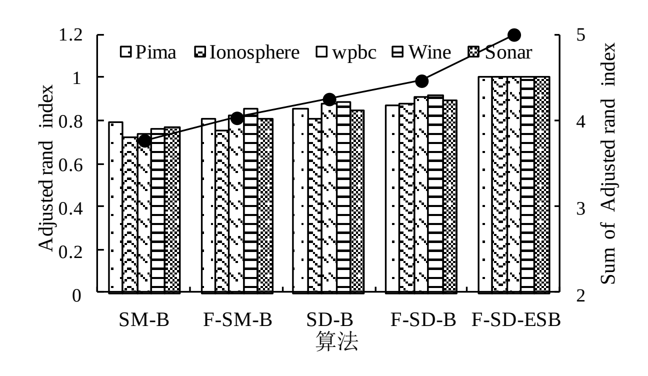


图 2　不平衡数据下 G-mean 的 Adjusted Rand Index 鲁棒性性能比较

由图 2 可知，本文所提的 Fisher-SDSMOTE ESBoostSVM 算法对各数据集的 b_m 均为 1，且具有最高的总和值，表明所提算法对不同空间结构及不同维度的数据不均衡分类问题均表现出良好的性能，在对比的其他四种算法中具有最好的鲁棒性。这是由于本文算法不仅考虑了样本的属性特性，利用改进的 SDSMOTE 算法使过采样更具针对性，而引入的“淘汰策略”对合成的正类样例进行二次筛选，提高了合成样本的质量，进而提高了模型的分类准确率。

3　算法在信用评分中的应用

3.1　数据准备和预处理

本文选用 UCI 数据集中 German 公开数据集，该数据集中包含 1 000 条贷款申请记录，其中 700 条为信誉良好的“good”客户，300 条存在违约情况的“bad”客户。数据集中每条记录对应 20 个变量描述其特征属性，其中定量数据类型定属为 13 个，包括现有账户状况、信用记录、信贷目的、储蓄账户、当前工龄、婚姻状况、其他应收账款、抵押类型、其他分期付款计划、住房情况、工作状态、电话状态、是否为外籍工作者；7 种数值属性包括持续时间，借贷额度、分期付款金额占可支配收入比例，现居住地居住时间，年龄、未清还贷款金额，赡养人数。

令 $X = (x_1, x_2, \cdots x_n)^T$ 代表信用评分参考信息变量，每个样例可表示为 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ni})^T$，全部样本可表示为 $S = \{(x_i, y_i)\}$，$i = 1,2,\cdots,N$；y_i 代表客户贷款偿还情况，$I = \{i | y_i = 1, i \in N, (x_i, y_i) \in S\}$ 表示信用良好的客户，$J = \{i | y_i = -1, i \in N, (x_i, y_i) \in S\}$ 表示信用较差的客户。因此，信用评分问题可以简单地描述成是否可以通过客户的特征属性 x_i 而准确地将他们分成优质与劣质客户[18]。

SVM 为基于距离度量的分类模型，其对数量间数量级差别比较敏感。为避免数量级差别对分类结果的影响，在模型训练实验前使用最大—最小规范化方法对数据进行规范化。

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

3.2　实验结果及性能分析

为验证信用评分问题中各评判指标的 Fisher 比率值对分类器的影响，首先对数据进行预处理，而后按照式（8）计算出各特征的 Fisher 比率值，并对这些特征以此进行降序排列，最后依次选取各个特征建立分类模型，计算模型的 G-mean 值和 F-measure 值。测试结果如图 3 所示。

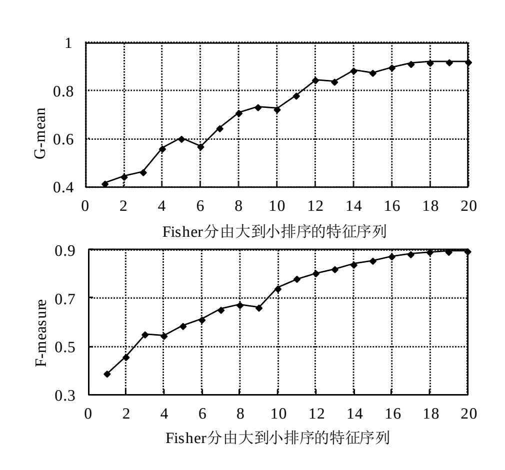


图 3　按 Fisher 分排序的特征分类模型测试结果

由图 3 可知，随着 Fisher 比率值的降低，相应特征对分类的影响逐渐减小，按照 Fisher 比率值大小排序后前 18 个特征对模型分类效果有较大影响，而其余特征对分类结果影响较小，可以忽略不计，故将其视为噪声特征删除。

实验过程以 MATLAB 2012b 为平台，采用十折交叉验证将数据集平均分成 10 份，训练数据集和测试数据集的比例为 1：9，每份数据集依次作为训练数据集。各类算法的设置亦同上。为避免机器学习不稳定性带来的随机影响，每折运行 10 次，实验次数共 100 次，最后得到每个评价指标的均值。同时将构建模型运行效果与 SMOTE-BoostSVM、Fisher-SMOTE-BoostSVM、SDSMOTE-BoostSVM、Fisher- SDSMOTE-BoostSVM 模型的实验结果对比，各评价指标计算结果如图 4 所示。

对比五种算法的实验结果可以看出，算法总体分类准确率从大到小的排名依次为 Fisher-SDSMOTE-ESBoostSVM、Fisher-SDSMOTE-BoostSVM、SDSMOTE-BoostSVM、Fisher -SMOTE-BoostSVM、SMOTE-BoostSVM，算法整体分类准确率随正类样本分类准确率的提高得到明显改进.另外，从图 3 中可以看出，

本文提出的 Fisher-SDSMOTE- ESBoostSVM 算法在整体分类准确率上有较大提高，说明本文所提出的改进过采样算法及“淘汰策略”通过产生新的质量较高的正类样本平衡训练信息，较好地解决了客户信用评分中的类别不平衡问题。综上实验结果表明，相比其他算法，本文提出的 Fisher-SDSMOTE-ESBoostSVM 算法在处理信用评分中类不平衡问题时效果较好，学习性能更优，具有较好的适用性。

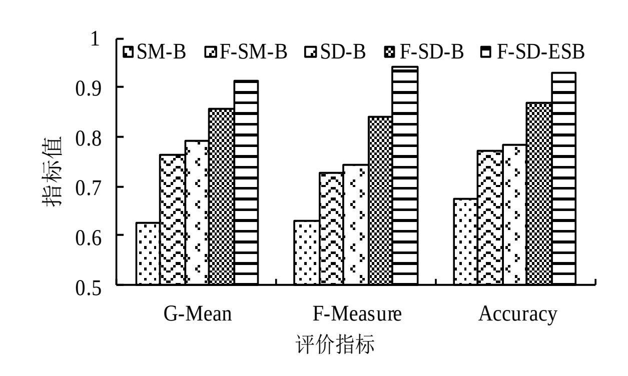


图 4 五种模型各指标值对比

3.3 与其他算法的对比

为进一步测试所构建算法的性能，本文将其与 KM-SMOTE-RF、Fisher-RUS-AdaboostSVM，Relief-CSCART 三种集成分类算法进行了对比。三种集成分类算法的集成策略见表 3。10 次实验结束后得到各评价指标均值如表 4 所示，各算法对应的 ROC 曲线如图 5 所示。

表 3 三种集成算法的集成策略描述

算法	策略
KM-Borderline-SMOTE-RF	K-Mean 聚类+基于边界的过采样算法 SMOTE+随机森林集成学习算法
Fisher-RUS-AdaboostSVM	Fisher 特征提取方法+欠采样算法 RUS+Adaboost +SVM 基分类器
Relief-CSCART	Relief 特征选择方法+代价敏感决策树

表 4 本文算法与其他三种集成学习算法对比情况

算法	G-mean	F-measure
KM-Borderline-SMOTE-RF	0.823	0.798
RUS-AdaboostSVM	0.882	0.845
Relief-CS-CART	0.895	0.873
Fisher-SDSMOTE-ESBoostSVM	0.923	0.896

从图 5 可以看出，四种算法对信用评分预测的表现是相当的，Fisher-RUS-AdaboostSVM、Relief-CS-CART 的分类效果相差不大，较优于 KM-SMOTE-RF 算法，说明了特征选择在处理不平衡数据分类问题上的有效性；Fisher-RUS-AdaboostSVM 算法更优于 Relief-CS-CART 算法，表明对于信用评分分类预测问题，从数据层面处理不平衡问题相对与从算法层面处理不平衡问题更具优势；而本文提出的 Fisher-SDSMOTE-ESBoostSVM 算法分类效果较优与其他三种算法，说明对过采样算法进行改进并结合“淘汰策略”确保合成样本的正确性使分类器对少数类样本具有了更强的学习能力，可以为类别不平衡信用评分问题提供一定的参考作用。

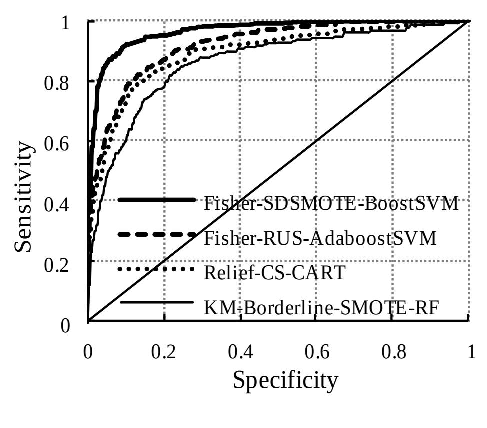


图 5 ROC 曲线对比

4 结束语

本文针对信贷行业中客户信用评分业务存在的类别不平衡问题，首先采用 Fisher 准则对信用评分中各评判指标进行选择，合理降低了数据维度，减小了数据间信息冗余；并提出一种基于支持度的 SDSMOTE 过采样算法指导正类样本的合成，有效避免了传统 SMOTE 算法合成样本的盲目性，使过采样更具针对性，进而提高正类样本的分类效率。以 Boosting 集成学习算法为框架， SVM 为基分类器，引入“淘汰策略”，删除基分类器中分类错误的正类样本，重新合成并更新样例权重，提高了合成样本的质量。

实验结果表明，本文所提出的算法相比其他集成算法具有较好的 F-measure 和 G-mean 值，应用到信贷行业客户信用评分预测中具有可行性和适用性。本文仅将客户分为信用好和信用差客户，如何对客户信用等级进行更详细的划分是今后研究的重点。

参考文献：

[1] 张婷婷. Logistic 回归及其相关方法在个人信用评分中的应用 [D]. 太原: 太原理工大学, 2017.

[2] 陆爱国, 王珏, 刘红卫. 基于改进的 SVM 学习算法及其在信用评分中的应用 [J]. 系统工程理论与实践, 2012, 32 (3): 515-521.

[3] 陈启伟, 王伟, 马迪, 等. 基于 Ext-GBDT 集成的类别不平衡信用评分模型 [J]. 计算机应用研究, 2018, 35 (2): 1-9

[4] Herrera F. On the use of map reduce for imbalanced big data using Random Forest [J]. Information Sciences An International Journal, 2014, 285 (3): 112-137.

[5] Blake C L, Merz C J. UCI Repository of machine learning databases [D]. Irvine (CA): University of California, 1998

[6] Chawla N V, Bowyer K W, Hall L O, *et al.* SMOTE: synthetic minority over-

sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16 (1): 321-357.

[7] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new oversampling method in imbalanced data sets learning [C]// Proc of International Conference on Intelligent Computing. 2005: 878-887.

[8] Nakamura M, Kajiwara Y, Otsuka A, *et al.* LVQ-SMOTE-learning vector quantization based synthetic minority over–sampling technique for biomedical data [J]. Biodata Mining, 2013, 6 (1): 16.

[9] 郭明玮，赵宇宙，项俊平，等. 基于支持向量机的目标检测算法综述 [J]. 控制与决策, 2014, 29 (2): 193-200.

[10] 徐乾，王文剑，张文浩. 处理非平衡数据的粒度 SVM 学习方法 [J]. 计算机工程与应用, 2011, 47 (24): 97-99+114.

[11] 李诒靖，郭海湘，李亚楠，等. 一种基于 Boosting 的集成学习算法在不均衡数据中的分类 [J]. 系统工程理论与实践, 2016, 36 (1): 189-199.

[12] 李雄飞，李军，董元方，等. 一种新的不平衡数据学习算法 PCBoost [J]. 计算机学报, 2012, 35 (2): 2202-2209.

[13] 黄海松，魏建安，康佩栋. 基于不平衡数据样本特性的新型过采样 SVM 分类算法 [J]. 控制与决策, 2017, : 1-10

[14] 赵清华，张艺豪，马建芬，等. 改进 SMOTE 的非平衡数据集分类算法研究 [J]. 计算机工程与应用, 2017, : 1-7.

[15] 周绍磊，廖剑，史贤俊. 基于 Fisher 准则和最大熵原理的 SVM 核参数选择方法 [J]. 控制与决策, 2014, 29 (11): 1991-1996.

[16] 古平，欧阳源遊. 基于混合采样的非平衡数据集分类研究 [J]. 计算机应用研究, 2015, 32 (2): 379-381+418.

[17] 陶新民，郝思媛，张冬雪，等. 基于样本特性欠取样的不均衡支持向量机 [J]. 控制与决策, 2013, 28 (7): 978-984.

[18] 韩璐，韩立岩. 正交支持向量机及其在信用评分中的应用 [J]. 管理工程学报, 2017, 31 (2): 128-136.